النماذج اللغوية الكبيرة العربية: تطوراتها، وتحدياتها وتطبيقاتها

جدول المحتويات

بسم الله الرحمن الرحيم

نظرة عامة

نظرة عامة:

يقدم هذا الفصل استعراضاً شاملاً ومعمقاً للنماذج اللغوية العربية في مجال الذكاء الاصطناعي ومعالجة اللغات الطبيعية. يبدأ الفصل بمقدمة تاريخية تتبع تطور هذا المجال، متبوعة بتقدمة اصطلاحية توضح المفاهيم والمصطلحات الأساسية للقارئ.

ينتقل الفصل بعد ذلك إلى صلب الموضوع، حيث يستعرض بالتفصيل ثلاثة أنواع رئيسية من النماذج اللغوية العربية: نماذج المفككات، نماذج المرمزات، ونماذج المرمزات الفاكّة. يتم شرح كل نوع بعمق، مع توضيح هيكليته، آلية عمله، ومميزاته. لتعزيز الفهم، يتم استخدام رسوم بيانية توضيحية لهذه النماذج.

يلي ذلك قسم مخصص لأهم التطبيقات العملية لهذه النماذج، مستعرضاً مجالات استخدامها المتنوعة وتأثيرها على مختلف القطاعات. كما يناقش الفصل التحديات الحالية التي تواجه تطوير وتطبيق هذه النماذج في السياق العربي، ويستشرف التوجهات المستقبلية في هذا المجال الحيوي.

يختتم الفصل بخلاصة تجمع أهم النقاط المطروحة، ويقدم قائمة شاملة بالمراجع العربية والأجنبية، مما يوفر للقراء مصادر قيمة لمزيد من التعمق والبحث في هذا الموضوع المتطور باستمرار.[1]

[1] يجدر الإشارة إلى أن نظام الإحالة اعتمد على طريقة (chicago) وقد استعملنا الأرقام للإحالة على المصادر فما كان بين قوسين [] فهو مصدر باللغة الانجليزية وما كان فيه حرف عين [ع] فهو مصدر باللغة العربية.

أبرز ما جاء في الفصل

النقاط الأساسية للفصل تنحصر في:

- تقديم عامة حول مجال المعالجة اللغوية والنماذج اللغوية الضخمة
- تعريف المصطلحات التقنية للمجال وتوطئة تاريخية مُختصرة.
- حصر ووصف النماذج اللغوية الكبيرة العربية بهيكَلِيَّاتها الثلاثة، وهي المفكك، والمرمز، والمرمّز الفّاك
- عرض أهم تطبيقات النماذج اللغوية الكبيرة العربية
- عرض التحديات العامة المتعلقة في المجال مما يُسْتَنتَجُ من البيانات المجموعة.
- مناقشة التوجهات البحثية المستقبلية.

أسئلة البحث

- ما هي الجهود المبذولة في تطوير النماذج العربية؟
- ما هي التطبيقات العملية التي تميزت بها هذه النماذج؟
- ما هي التحديات الرئيسية التي ظهرت في هذا المجال؟

1. مقدمة تاريخية واصطلاحية

تستعرض هذه المقدمة المصطلحات الأساسية المستخدمة في هذا البحث، حيث تم الاعتماد في ترجمتها على معجم سدايا للبيانات والذكاء الاصطناعي [ع6]. ويتضمن البحث عرضًا موجزًا للتطور التاريخي في مجال المعالجة اللغوية، بدءًا من مرحلة المعالجة اليدوية وصولاً إلى تطوير نماذج المحولات (Transformers) التي أحدثت تحولاً جوهريًا في المجال، مما أدى إلى ظهور النماذج اللغوية المتقدمة.

كما يتطرق البحث إلى المسار التاريخي لتطور معالجة اللغة العربية، مع التركيز على أبرز المحطات التطويرية في هذا السياق. وتجدر الإشارة إلى أن التفاصيل التاريخية الشاملة سيتم مناقشتها في فصل مستقل، لذا سيقتصر العرض هنا على الإشارة الموجزة لأهم المراحل التطويرية، مع التركيز بشكل خاص على الجوانب المتعلقة بهيكلية المحولات. [61][62][ع8]

1.1. التوطئة التاريخية

- في المرحلة الأولى، والتي امتدت بين عامي (1950-1960)، اعتمدت أنظمة المعالجة اللغوية على القواعد المُهندَسة يدوياً (Hand-crafted Rules). وتنقسم هذه الأنظمة إلى قسمين رئيسيين: المحللات النحوية (Syntactic Parsers)، والمسترجعات القائمة على القواميس (Dictionary-based Retrievers). من أبرز أمثلة القسم الأول نظام "إلزا" (ELIZA) للمحادثة الآلية، الذي اعتمد على التحليل اللغوي السطحي باستخدام التعابير النمطية (Regular Expressions). أما في القسم الثاني، فيُمكن الإشارة إلى أنظمة التحليل المشاعري (Sentiment Analysis) القائمة على قوائم الكلمات المحددة مسبقاً. وفي سياق اللغة العربية، التي شهدت تأخراً نسبياً في هذا المجال، ظهرت بعض الأنظمة المتخصصة في التحليل الصرفي والاشتقاقي. وتجدر الإشارة إلى أن هذه الفترة شهدت أيضاً ظهور أول الأبحاث في مجال الهياكل العصبية من خلال نموذج "المُدرِك" (Perceptron) على يد العالِم "روزنبلات" (Rosenblatt).

  في المرحلة الثانية، والتي امتدت من الثمانينيات إلى أواخر التسعينيات (1980-1990)، شهد المجال تطوراً نوعياً مع ظهور الطرق الإحصائية، وخاصة نموذج ماركوف الخفي (Hidden Markov Model - HMM). وقد أتاح هذا التطور ظهور تطبيقات أكثر تعقيداً، مثل وسم أقسام الكلام (Part of Speech Tagging - POS)، والتعرف على الكيانات المُسماة (Named Entity Recognition - NER). وفيما يخص اللغة العربية، شكّل تأسيس شركة صخر [ع7] نقطة تحول مهمة، حيث أسهمت في تطوير العديد من التطبيقات اللغوية. كما شهدت هذه الفترة إجراء دراسات إحصائية متنوعة، شملت:

  - تحليل المعاجم العربية الرئيسية مثل "الصحاح" و"تاج العروس"
  - تطوير أنظمة التشكيل الآلي (Diacritization Systems)
  - إنشاء أنظمة الترجمة الآلية (Machine Translation Systems)
  - إجراء دراسات إحصائية متخصصة على النص القرآني

- في المرحلة الثالثة، التي امتدت من التسعينات إلى بداية الألفية الثانية (1990-2000)، ظهر توجه يتقارب مع المرحلة الأولى في اعتماده على القواعد الجاهزة (Rule-based Systems)، إلا أنه تميز بعمق تطبيقي مختلف. حيث انصب التركيز على تطوير معاجم متخصصة (Specialized Lexicons)، والتي وُظِّفت في التصنيف المتخصص للمدونات النصية (Specialized Corpus Classification). وعلى صعيد اللغة العربية، واستمراراً لإنجازات المرحلة الثانية، تم إنجاز تطوير معجم تركيبي شامل يجمع بين التوليد اللغوي (Language Generation)، التحليل الصرفي (Morphological Analysis) والتحليل النحوي (Syntactic Analysis). كما شهدت نهاية هذه المرحلة وبداية المرحلة التالية تحولاً مهماً في مجال معالجة اللغة العربية، تجلى في عدة جوانب:
    - تزايد الاهتمام بمعالجة اللغة العربية نتيجة للتطورات السياسية (أحداث سبتمبر/أيلول)
    - إطلاق مشروع "بنك العربية الشجري" (Penn Arabic Treebank - PATB)
    - ظهور حاجة ملحة لتطوير وتجميع مدونات عربية ضخمة (Large Arabic Corpora)

- المرحلة الرابعة، والتي تمتد بين عامي 2000 و2017، تتميز بظهور نهج المعالجة التضمينية (Embedding Processing) أو ما يُعرف بتقنيات التمثيل المتجهي (Vector Representation). في هذه المرحلة، شهدنا تحولًا جوهريًا في تمثيل النصوص والكلمات، حيث أصبحت تُجسّد على هيئة متجهات رقمية (Digital Vectors) متعددة الأبعاد. هذه المتجهات تتخذ شكلين رئيسيين: إما متجهات عدِّية (Sparse Vectors) أو متجهات كثيفة (Dense Vectors). أدى هذا التطور إلى توسيع نطاق تطبيقات معالجة اللغة الطبيعية، حيث وُظّفت هذه التقنيات في مجموعة واسعة من المهام اللغوية، بما في ذلك تصنيف النصوص واسترجاع المعلومات، فضلًا عن تطبيقات أخرى متنوعة. من بين الأساليب البارزة في هذا المجال، نجد تقنية TF-IDF (تكرار الكلمة - التكرار العكسي للوثيقة)، إلى جانب نماذج التضمين المتقدمة مثل Word2Vec. جدير بالذكر أن نموذج Word2Vec ومشتقاته قد حظي باهتمام بالغ في مجال الأبحاث اللغوية العربية، حيث شهدنا توظيفًا مكثفًا لهذه التقنيات في العديد من الدراسات والتطبيقات العملية. وقد أشارت الأبحاث [59,60] إلى الأثر الملموس لهذه التقنيات في تعزيز فهم وتحليل اللغة العربية آليًا.
كما شهدت هذه المرحلة، وتحديدًا حوالي عام 2012، تطورًا ملحوظًا في مجال الشبكات العصبية العميقة. برزت الشبكات العصبية الالتفافية (CNN) في مجال معالجة الصور، بينما أظهرت الشبكات العصبية المتكررة (RNN) قدرات واعدة في معالجة تسلسلات البيانات، بما في ذلك النصوص. هذه التقنيات مهدت الطريق لتطورات أكثر تقدمًا في المرحلة اللاحقة.

- تمتد المرحلة الخامسة في تطور معالجة اللغة الطبيعية من عام 2018 حتى وقتنا الحاضر، وتتميز بالتقدم الملحوظ في تقنية المحولات (Transformers). هذه التقنية، التي قُدمت أولاً في عام 2017 ، أحدثت ثورة في مجال معالجة اللغات الطبيعية بفضل قدرتها الفائقة على معالجة العلاقات طويلة المدى في النصوص. في عام 2018، شهد المجال نقلة نوعية مع ظهور نموذج BERT (Bidirectional Encoder Representations from Transformers).
يعتمد هذا النموذج على هيكلية المُرمِّز ثنائي الاتجاه، مما مكّنه من إحداث طفرة في مهام الفهم اللغوي، وخاصة في مجالات تصنيف النصوص واستخراج المعلومات والإجابة عن الأسئلة. في العام التالي، تم تقديم نموذج GPT-2 Generative Pre-trained Transformer 2، الذي أظهر قدرات متقدمة في توليد النصوص. هذا النموذج شكّل الأساس لتطوير سلسلة نماذج GPT اللاحقة، بما في ذلك GPT-3 وGPT-3.5، والتي بدورها مهدت الطريق لظهور

تطبيق ChatGPT في نهاية عام 2022. هذا التطبيق حظي بشهرة عالمية واسعة، مسلطًا الضوء على إمكانات الذكاء الاصطناعي في التفاعل اللغوي شبه البشري. جدير بالذكر أن هذه الفترة شهدت أيضًا تطورات مهمة في معالجة اللغات الأخرى غير الإنجليزية، بما في ذلك اللغة العربية. تم تطوير نماذج خاصة باللغة العربية مثل AraBERT وCAMeLBERT ، والتي أظهرت أداءً متميزًا في مختلف المهام اللغوية العربية.

يجدر بنا أن نستعرض، ولو بشكل مختصر، السبب الرئيسي وراء الانتشار الكاسح لنماذج المحولات في مجال المعالجة اللغوية خاصة، ومنه إلى مختلف المجالات عموماً. كما أشرنا من قبل، بدأت القصة عام 1958 مع صدور أول بحث عن العصبونات الحوسبية باسم "المُدرك (Perceptron)"، لكن لأسباب تاريخية وبحثية وحوسبية متعددة، لم تبدأ تطبيقات النماذج العصبية بالظهور بوضوح إلا في عام 2012. شهدنا حينها انتقالاً جوهرياً من الخصائص المستخرجة يدوياً إلى تلك المتعلمة آلياً في الشبكات العصبية، والتي تطورت لتصبح ذات طبقات عديدة.

رغم هذا التقدم، عجزت هذه النماذج العصبية التقليدية عن فهم السياقات وتتابعها بشكل كافٍ لاستخدامها بفعالية في المعالجة اللغوية الآلية توليداً وفهماً، و إن لم تخل الساحة من محاولات متفرقة.

ثم ظهرت هيكلية الشبكة العصبية التكرارية (RNN)، والتي عالجت جزئياً مشكلة تذكر السياقات اللغوية باستخدام الحالات الخفية. تحسن أداء التطبيقات اللغوية مع هذه الهيكلية، خاصة بعد تطويرها إلى "الذاكرة القصيرة المدى المُطولة" (LSTM). ومع ذلك، ظلت هناك تحديات في التعامل مع السياقات الطويلة وبطء عملية التدريب.

نتيجة لذلك، برزت الحاجة إلى هيكلية تجمع بين سرعة التدريب (معتمدة على التوازي بشكل أساسي) وفهم السياق بشكل أعمق. هنا ظهر مبدأ "الانتباه (Attention)"، الذي يسمح للأجزاء اللغوية بالتواصل فيما بينها، حيث يحصل كل جزء على المعلومات التي يحتاجها من محيطه، إعطاءً وأخذاً.

على أساس هذا المبدأ، تم بناء نماذج المحولات. وهنا تجدر الإشارة إلى أن المحولات تمثل عودة إلى الشبكات ذات التغذية الأمامية (FFN)، مع فارق جوهري يتمثل في حقن السياق بأكمله دفعة واحدة، مع ترميز المواقع لضمان فهم تسلسل المدخلات. هذا النهج، إلى جانب التطور الكبير في القوة الحوسبية، مكّن هذه الهيكلية من التفوق على جميع النماذج السابقة.

منذ ظهورها في عام 2017 وحتى يومنا هذا، لا تزال نماذج المحولات في تطور وتحسين مستمر، مما يعزز قدرتها على فهم وتوليد اللغة بشكل أكثر دقة وفعالية.

1.2. التقدمة الاصطلاحية

نذكر فيها أهم المصطلحات المُستخدمة في الفصل:

- المقسم (Tokenizer) وهو أداة تُستخدم لتقسيم النصوص المكتوبة بناءً على مجموعة أجزاء لغوية محدودة (مميزة) تم بناءها مُسبقًا
- الجزء اللغوي (Token) وهو قطعة نصية (قد تكون أقل من كلمة أو أكثر أو مساوية) تم استخراجها باستعمال أحد خوارزميات الضغط النصي لتكوين مجموعة محدودة من الأجزاء التي يُمكنُ تمثيل المدونات النصية بها.
- خوارزمية (BPE) وهي أحد أشهر الخوارزميات المُستعملة في ضغط النصوص ويكثر استعمالها في النماذج اللغوية. وهذا الاختصار (BPE) يعني (Byte Pair Encoding) أي ترميز أزواج البايت.
- هيكلية المحولات (Transformers): وهي أحد أشهر وأحدث بُنى الشبكات العصبية الصادرة عام 2017 وأول ما صدرت كانت لمهمة الترجمة مكوَّنةً من جُزئين الأول مُرَمِّزْ (مشفر) والثاني مفكك الترميز.
- المحولات المرمِّزة (Encoder): هيكلية تستعمل في تدريب النموذج جزءَ المرمز فقد وتُستعمل عادة في مهام الفهم اللغوي.
- المحولات المفككة (Decoder) : وهذه تستعمل فقط جزءَ المفكك من المحولات وتُستعمل عادة في توليد النصوص.
- المعامل (Parameter) : رقم متغير يتغير أثناء عملية تدريب نموذج الشبكة العصبية (بغض النظر عن بُنيتها) يتعلم بتغييره النموذج النمط المخفي بين المُدخلات والمُخرجات.
- المدونة (Corpus\Dataset): مجموعة نصية وفي سياق النماذج اللغوية تكون خامًا غير مُوسَّمة تُستعمل في تعليم النموذج لغةً معينة.
- النموذج اللغوي الضخم (Large Language Model): مُصطلح يشير إلى نموذج ذي عدد كبير من المعاملات وفي الفترة الأخيرة أُصطلِحَ على ما تجاوز مليار مُعامل.
- المحسِّن (Optimizer): خوارزمية تُستخدم في تحديث معاملات النموذج وتغييرها أثناء عملية التدريب.
- التضمين (Embedding): متجه بأرقام حقيقية يُستعمل في سياق النماذج اللغوية لتمثيل مركب نصي سواء كان كلمةً أو أكثر، يُعَبِّرُ عن دلالة النص في الفضاء الجبري (المُتَّجهي).
- خوارزمية (PPO): أحد خوارزميات التعلم التعزيزي لتحسين أداء الوكلاء في المهمات الصعبة (وفي حالة النماذج اللغوية أُستعملت في تحذية أو تنسيق (Alignment) النص المكتوب بحسب قواعد معينة عادة تكون لمنع الردود المسيئة أو الضارة).
- نافذة السياق (Context window): رقم ثابت من الأجزاء اللغوية يمر على النموذج أثناء تعلمه فيفهم منه سياقا محددا.

2. النماذج اللغوية العربية

في هذا القسم سنتكلم، عن النماذج اللغوية بهيكليَّاتها الثلاث بحيث نعرض كل هيكلية بتعريف مُختصر ثم نسرد أهم النماذج لكل واحدة ونعرض اسم النموذج والجهة والدولة وعدد المعاملات وبيئة التدريب وحجم البيانات (بأعداد الأجزاء اللغوية) والتدريب ومصدر البيانات.

2.1. الباب الأول في نماذج المفككات

نماذج المفككات أصبحت الهيكلية الأبرز في معالجة اللغات الطبيعية خلال السنوات الأخيرة، مع تطبيقات شهيرة مثل **ChatGPT** معتمدة عليها. هذه النماذج التوليدية تعمل بآلية بسيطة وهي أن تأخذ نافذة سياق محددة، ثم توقع الأجزاء اللغوية التالية تباعاً. رغم بساطتها، أظهرت هذه النماذج - عند تدريبها على نصوص ضخمة وبعدد هائل من المعاملات - فهماً لغوياً متقدماً. بعد الضبط الدقيق، أصبحت قابلة للاستخدام في مجموعة واسعة من التطبيقات اللغوية كالتلخيص، الإجابة عن الأسئلة، والترجمة. نجاح هذه النماذج يعود لحجمها الكبير، كمية البيانات الضخمة المستخدمة في تدريبها، وتقنيات التدريب المتقدمة، مما جعلها أداة قوية وشائعة في مجال الذكاء الاصطناعي ومعالجة اللغات الطبيعية.

نستعرض فيما يلي أهم النماذج اللغوية المفككة للغلة العربية.

جدول 1: جدول النماذج المفككة

العائلة	الاسم	تاريخها	المصدر	عدد المعاملات	حجم البيانات[2]	معلومات المقسم	بيئة التدريب	مصدر البيانات	معلومات التدريب
عائلة AraGPT[3] [1]	AraGPT2-base	2020	الجامعة الأمريكية في بيروت	135 مليون	77 جيجابايت	مقسم BPE وعدد وحداته اللغوية 64000،	باستعمال وحدة معالجة تنسرية واحدة TPU بمكتبة gpt-2-simple[4]	مقالات عربية فصيحة وفيها أيضا نصوص مسحوبة من الشبكة و ويكيبيديا (الموسوعة الحرَّ)	تم التدريب على وحدة معالجة تنسر TPU. مواصفات النموذج: طول سياق 1024 وبُعد التضمين 768 وله 12 رأسَ انتباه و12 طبقة ويستخدم مُحَسِّن LAMB
	AraGPT2-medium			370 مليون					نافذة سياق 1024 وحدة لغوية، بُعد تضمين 1024، 16 رأس انتباه، 24 طبقة، محسّن LAMB،

[2] يُذكر عدد الأجزاء اللغوية (بما ذكرته الورقة باستعمال مُقَسِّمهم) و إلا فيُذكر حجم البيانات.

[3] الرابط: https://huggingface.co/aubmindlab

[4] وهذا رابطها https://github.com/minimaxir/gpt-2-simple

	AraGPT2-large			792 مليون			باستعمال وحدة معالجة تنسرية واحدة TPU بمكتبة gpt2-ml[5].		نافذة سياق 1024 وحدة لغوية، بُعد تضمين 1280، 20 رأس انتباه، 36 طبقة، محسّن Adafactor،
	AraGPT2-mega			1.46 مليار					نافذة سياق 1024 وحدة لغوية، بُعد تضمين 1536، 25 رأس انتباه، 48 طبقة، محسّن Adafactor،
عائلة Bloom[6] [2]	BLOOM-176B (وهناك خمس نسخ ثانوية)	2022	شركة بيج ساينس BigScience بالتعاون مع جهات أخرى.	176 مليار	1.6 تيرابايت والعربية منها 75 جيجابايت	BPE - 250,680 وحدة لغوية	-	مدونة "مصادر" و OSCAR [3]	نافذة سياق 2048 وحدة لغوية، بُعد تضمين 14336، 70 طبقة، 112 رأس انتباه، التضمينات الموضعية بخوارزمية ALIBI، أضيف 200 جزء لغوي للاستخدام المستقبلي.
عائلة[7] ياسمين (Jasmine) [5]	Jasmine-350M	2022	جامعة بريتش كولومبيا British Columbia وجامعة محمد بن	350 مليون	244 جيجابايت	BPE 64,000 وحدة لغوية	استعمال مكتبة GPT-Neo	نصوص من الشبكة، مقالات إخبارية، كتب، ويكيبيديا	12 طبقة، 12 رأس انتباه، بُعد تضمين 768،
	Jasmine-1.3B			1.3 مليار					24 طبقة، 16 رأس انتباه، بُعد تضمين 2048،

[5] وهذا رابطها https://github.com/imcaspar/gpt2-ml/

[6] الرابط: https://huggingface.co/bigscience

[7] الرابط: https://huggingface.co/UBC-NLP/Jasmine-350M

	Jasmine-2.7B		زايد MBZUAI	2.7 مليار					32 طبقة، 32 رأس انتباه، بُعد تضمين 2560.
	Jasmine-6.7B			6.7 مليار					32 طبقة، 32 رأس انتباه، بُعد تضمين 4096.
عائلة[8] AceGPT [6]	AceGPT-7B	2023	معهد شنجن Shenzhen الصيني للبيانات الضخمة و جامعة هونج كونج في جنشن Shenzhen وجامعة كاوست في السعودية	7 مليار	30 مليار جزء لغوي 19 مليار جزء عربي و11 مليار جزء	-	-	بيانات عربية نسخة ArabicText 2022	اعتمدوا على نموذج LLAMA-2 بأن بدؤوا تدريبه على اللغة العربية. -
	AceGPT-13B			13 مليار	10 مليار جزء عربي	-	-		
AceGPT 2	AceGPT 2-7B	2024		7 مليار	30 مليار جزء لغوي	IBPE [9]	2368 كرت شاشة		
	AceGPT 2-13B			13 مليار					
جيس[9] Jais (الجيل الأول) [10]	Jais-13B	2023	شركة Inception، جامعة محمد بن زايد، أنظمة Cerebras	13 مليار	72 مليار جزء لغوي (18 مليار عربي)	BPE (84,992 جزءًا)	الحاسوب الخارق Condor Galaxy	18 مليار جزء عربي مترجم (3 مليار من ويكيبيديا و15 مليار من الكتب)، باقي البيانات من المدونات والنصوص العربية، 232	النموذج يحتوي على 40 طبقة، 40 رأس انتباه، بُعد التضمين 5120، استُخدمت متجهات ALiBi ودالة التنشيط SwiGLU، نافذة سياق بطول 2048، محسن AdamW،

[8] الرابط: https://huggingface.co/FreedomIntelligence

[9] على الرابط الجِيْلَان: https://huggingface.co/inceptionai

								مليار جزء إنجليزي (46 مليار منها نص برمجي من GitHub)	
	Jais-30B			30 مليار					التركيز على مسائل الأمان (العنصرية، تسريب المعلومات، الأذية، الجريمة، التلاعب النفسي).
Jais-Family [11]	Jais-Family (30B)	2024		30 مليار				بيانات عربية (490 مليار جزء) مشابهة للنموذج الأول ويُنتبه إلى أنّ هذه الأجزاء لأكبرها وإلا فالصغار أقل.	تم تدريب هذه النماذج (من هذه العائلة) على مجموعة بيانات جديدة بحجم 490 مليار جزء لغوي عربي، نافذة سياق بلغت 16,000 جزء، مع تحسين الأداء على المدونات اللغوية العامة، وتم تقييم النماذج على نفس أسس تقييم الجيل الأول مع تحسينات في الأداء.
	Jais-Adapted (LLaMA-2)			70 مليار	320 مليار جزء لغوي	IBPE (32,000 جزء جديد)			ضبط دقيق لنموذج لاما-2 بمفاتيح خاصة باللغة العربية باستخدام 32,000 جزء لغوي جديد، مع تدريب ضبطي يشبه عملية AceGPT، تم

									تقييم النتائج على مدونات مترجمة تم مراجعتها من قبل مراجعين عرب، كما تم التركيز على جودة النصوص الناتجة.
نون[10]	نموذج نون	2024	شركة نون، بالتعاون مع شركة نسيج السعودية	-	-	BPE (250,880 جزء لغوي)	-	-	يحتوي النموذج على بُعد تضمينات 4096، 32 رأس انتباه، 30 طبقة، نافذة سياق بطول 2048 جزء لغوي، واستخدام متجهات ALiBi. لم تُنشر تفاصيل حول بيانات التدريب أو منهجية التقييم، لكن النموذج متاح على الشبكة دون تقرير رسمي.
عائلة[11] Aya-23 [12]	Aya-23-8B	2024	شركة كوهير الأمريكية Cohere	8 مليار	-	BPE (256,000 جزء لغوي)	إطار Fax المعتمد على JAX	النصوص المولدة والمترجمة، 200 ألف مثال (للضبط الدقيق) [13]	بُعد التضمين 4096، 32 طبقة، 32 رأس انتباه، استخدم Grouped Query Attention،
عائلة[11] Aya-23 [12]	Aya-23-35B	2024	شركة كوهير الأمريكية Cohere	35 مليار	-	BPE (256,000 جزء لغوي)	إطار Fax المعتمد على JAX	النصوص المولدة والمترجمة، 200 ألف مثال (للضبط الدقيق) [13]	بُعد التضمين 8012، 40 طبقة، 64 رأس انتباه

[10] الرابط: https://huggingface.co/Naseej/noon-7b

[11] الرابط: https://docs.cohere.com/docs/chat-api وهكذا نزلت عائلة جديدة اسمها "Aya Expanse" بنفس الحجم وكذلك عديدة اللغات وهي تعتبر تحسينا للعائلة الحالية.

عائلة[12] Arabian GPT [15]	ArabianGPT-134 M	2024	معمل الروبوتات وإنترنت الأشياء في جامعة الأمير سلطان في السعودية	134 مليون	1.8 مليار جزء لغوي	sentence piece (64,000)	استعمُلت مكتبة Hugging Face	-	نافذة سياق 768 جزء لغوي. (والباقي نفسُهُ في AraGPT)
	ArabianGPT-345 M			345 مليون	3.34 مليار جزء لغوي			-	-
عائلة علام ALLAM [16]	ALLAM-7B نموذج 7 مليار (إصدار دُرِّبَ من الصفر)	2024	الهيئة السعودية للبيانات والذكاء الاصطناعي (سدايا SDAIA) والمركز الوطني للذكاء الاصطناعي (SCAI)	7 مليار	540 مليار جزء لغوي عربي (نصفها مترجم، نصفها مجموع)	مقسم لاما-2 مع تعديلات	إطار Megatron دُربَ على 1024 كرت A100	النصوص العربية الأصلية من الكتب والأخبار وبيانات الشبكة وأما المترجمة فهي كتب وبعض النصوص المنشورة	
	نموذج 7 مليار (نسخة من لاما-2)			7 مليار					نفس مواصفات التدريب والبيانات، لكن النموذج هو نسخة مُعدلة من لاما-2.
	نموذج 13 مليار (نسخة من لاما-2)			13 مليار					
	نموذج 70 مليار (نسخة من لاما-2)			70 مليار					
سلمى[13] SILMA	نموذج واحد	2024	شركة سلمى المغربية	8 مليار	-	مقسم جيما-2 Gemma-2	-	-	أحدث نموذج عربي صدر حتى الآن، أصله نموذج جيما-2 من جوجل، تم تدريبه على نصوص عربية، وليس هناك تفاصيل عنه

[12] الرابط: https://github.com/riotu-lab/aranizer

[13] الرابط: https://huggingface.co/silma-ai/SILMA-9B-Instruct-v1.0

عائلة[14] QWEN2 [17]	QWEN2-72B	2024	شركة علي بابا الصينية	72 مليار	7 ترليار جزء لغوي (غير معروف نسبة العربية فيها)	BPE 151,643 جزء لغوي	-	-	نموذج أجنبي ولكنّه جيد جدا في اللغة العربية وصدر له أيضا إصدار قريب باسم QWEN2.5 وهو فرد من عائلة بأربع أفراد.
عائلة Command R+	Command R+	2024	شركة كوهير Cohere	104 مليار	عشر لغات منها العربية (وغير معروف عدد الأجزاء اللغوية)	-	-	-	نموذج متعدد اللغات (عشر لغات) منها الإنجليزية والعربية والصينية وغيرها طول السياق 128 ألف جزء لغوي. وهناك نماذج أخرى في العائلة أصغر منها 7 مليارات.
-	نموذج فنار[15]	2024	الجهة الراعية: وزارة الاتصالات وتكنولوجيا المعلومات القطرية (وعدد من الجهات الأخرى)	7 مليارات	لعل التدريب كان على ترليار جزء لغوي.	-	-	اللغة العربية ولغات أخر[16]	تم تدريب النموذج على 220 كرت شاشة.

[14] الرابط: https://huggingface.co/Qwen وأثناء كتابة هذه الأوراق صدرت نسخة جديدة 2.5 منه وهي على نفس الرابط.

[15] موقع فنار: https://fanar.qa/

[16] ليس هناك تقرير تقني عن النموذج ولكن من تجربته لعله يعرف 13 عشرَ لغة وقد استفيد من مقابلة مع الدكتور محمد طباخ في بعض معلومات النموذج.

-	نموذج عقل AQL[17]	2024	جهة wideBot	7 مليارات	-	-	-	العربية ولغات أخرى	-

وآخر ما نتعرض إليه في عرض النماذج المفككة هي نماذج الشركات الضخمة والتي لا علمَ لنا إلا باسمها وتقييماتها وهي نماذج شركة OpenAI[18] وشركة Google[19] وشركة [20](Claude) وكلها تدعم اللغة العربية ولكن لا معلومات لدينا على البيانات التي تدربت عليها ولا حجمها ولا طُرق تدريبها لأنها مغلقة المصدر تجارية التوجه وهي للآن تُعتبر أكثر النماذج مُعتمدًا عليها واستعمالا (خاصة ChatGPT) إذ مما يلاحظ من خلاصات الأوراق السابقة أن أداء النماذج العربية المدربة قد يتجاوزها في بعض المهمات بطفيف ولكن عند الاستخدام التطبيقي فالغالب يستعملها نظرًا لوجود إطار برمجي API سهل يُوفِّرُ عناء تنزيل النماذج الضخمة وكلفة تشغيلها من بين عتاد وخبرة. ولكن لا يُنكر أيضا وجود أُطُر برمجية لبعض النماذج السابقة مثل نماذج شركة كوهير[21] وعلام (ولكنَّه محدود جدا على خوادم IBM) وكذا فإن لنموذج جيس واجهة استخدام يُسْتَجرب منها[22]. وهنا إشارة إلى أن التعامل اللغوي في النماذج اللغوية غالبه (بل قل كله) متعامل مع اللغات اللاتينية (وهناك جهود أخرى في اللغة الصينية ولكنها غير مفيدة للعربية لاختلاف الأنظمة) من اليسار إلى اليمين اتجاها ولغة ولا بُدَّ من دراسة دقيقة للمقسمات الموجودة وتأثير ذلك على اللغة العربية ومحاولة تحسينها أو حتى تطوير جديدٍ ليتناسب مع طبيعة اللغة العربية، وهناك بعض الدراسات مثل [64] التي قارنت مقسمات متنوعة على مهمة التصنيف وأخرى على أثر المقسمات على أداء النموذج المرمز [67]

---

[17] على الرابط التالي: https://widebot.ai/introducing-aql/

[18] وهو المعروف بتطبيق (ChatGPT)

[19] وهو المعروف باسم (GEMINI) وقديما كان اسمه (BARD)

[20] وهو أقل النماذج شهرة مقارنة بسابقيه واسمه (Sonnet 3.5) وأداؤه العربي جيد.

[21] على موقعهم: https://docs.cohere.com/

[22] على الرابط: https://jais.inceptionai.ai/c/

2.2. الباب الثاني في نماذج المرمزات

نماذج المرمزات هي النماذج المتعلقة بمهمات الفهم اللغوي (عادة التصنيف بأنواعه العامّة أو المخصصة)، بحيث يُعطى النموذج نص بنافذة سياق معينة ثم يقوم بتوقع أجزاء لغويّة تم اختيارها عشوائيا من النص المعطى، وهذا يُعطى النموذج فهمًا دلاليا وسياقيا. ويُضاف أيضا أن هذه النماذج تُستعمل في التمثيل الرقمي للنصوص (بأن يُحوِّل النص إلى متجه يُمثِّله معنويا) وسنشير إلى نماذج المحولات الجُمَليّة Sentence Transformers وأشهر العربيات منها.

جدول 2: جدول النماذج المرمِّزة

الاسم	تاريخها	المصدر	عدد المعاملات	حجم البيانات[23]	معلومات المقسم	بيئة التدريب	مصدر البيانات	معلومات التدريب
ArBERT [18]	2020	جامعة بيروت الأمريكية	أربع نسخ 136 مليون إلى 1.5 مليار	77 جيجابايت (200 مليون جملة)	مقسم BertWordpiece	TensorFlow Research Cloud (TFRC)	المقالات الإخبارية	تم تدريب النماذج على كرت المعالجة التنسرية TPU. وأيضا فقد صدرت له نسخة أخرى وليس لها ورقة بحثية[24]
MARBERT [19]	2021	جامعة بيريتش كولومبيا British Columbia	163 مليون	6.5 مليار جزء	100,000 ألف جزء مميز في مقسم WordPiece	(TFRC). خدمات جوجل السحابية باستعمال	الكتب والمقالات الإخبارية ونصوص من الشبكة	معالج تنسري TPU. وتم التدريب بعدد 42 دورة.
ARBERT [19]				15.6 مليار جزء			مليار تغريدة من تويتر	أُستعمل معالج تنسري TPU. وتم التدريب بعدد 36 دورة. على مدى 40 يومًا. وقد صدرت نُسخة جديدة[25]

[23] يُذكر عدد الأجزاء اللغوية (بما ذكرته الورقة باستعمال مُقَسِّمهم) و إلا فيُذكر حجم البيانات.

[24] ليس لها ورقة بحثية و إنما نُشرت فقط: https://huggingface.co/aubmindlab/bert-base-arabertv02

[25] أيضا ليس للنسخة الجديدة ورقة و إنما نشرت فقط: https://huggingface.co/UBC-NLP/ARBERTv2

GigaBERT [20]	2020	جامعة أوهايو الأمريكية، ومعهد التقنية في جورجيا أمريكا	125 مليون	10.4 مليار جزء (4.3 مليار منها عربي)	مقسم wordPie ce فيه 50,000 جزء مميز (باللغتين)	-	مدونة GigaWork وويكيبيديا ونصوص من الشبكة	استُعملت البرمجيات الأصلية لتدريب BERT في مكتبة TENSORFLOW مع محسن آدم.
JABER [21]	2021	شركة هواوي وجامعة مونتريال في كندا	135 مليون	115 جيجابايت	مقسم BBPE بعدد أجزاء 64,000	16 خادم من شركة هواوي بعدد 8×16 كرت شاشة V100 بـ32 جيجابايت	الأخبار وويكيبيديا ونصوص الشبكة (مُعالجة)	نفس النصوص البرمجية لتدريب BERT بعدد 15 لفة باستعمال محسن آدم AdamW بعدد 16 ساعة
SABER [21]			365 مليون					عدد 5 لفات باستعمال محسن آدم AdamW بعدد 32 ساعة
QARIB [63]	2021	معهد قطر للبحوث الحوسبية، ومجمع بحوث HBKU	160 مليون	تجارب كثيرة أكبرها 60 جيجابايت (7.6 مليار)	BPE	Google Cloud (TPU (v2.	مقالات اخبارية وتغريدات مُنتقاة	تم التدريب على موقع كولاب COLAB باستعمال كرت المعالجة التنسرية TPU.

وقبل الانتقال للنماذج المرمزة الفاكَّة تجدر الإشارة لنماذج (هي في الأصل مرمزة) مهمتها استخراج المعاني الدلالية (اللغوية و المعنوية) من النصوص وتمثيلها على هيئة متجهات ليسهل البحث وترتيبها وفهرستها وغيرها من التطبيقات الكثير (وسيشار لها في قسم التطبيقات) وسنذكر هنا أهم العوائل ولن نطيل بذكر كل تفاصيلها لأنها كثيرة جدا.

جدول 3: النماذج المُتْجِهَة

اسم العائلة	المصدر	وصف مختصر
عائلة[26] SentenceTransformers	مكتبة SentenceTransformers	هذه المكتبة تحوي عددًا كبيرا من النماذج المُتجهة عديدة اللغات (ومنها العربية) مثل نماذج: **MiniLM, mpnet, LaBSE** وقد استعملت في عدد كبير من الأوراق البحثية
عائلة Matryoshka[27] [66]	معمل الروبوتات و إنترنت الأشياء في جامعة الأمير سلطان	طريقة جديدة تعتمد على النماذج المرمزة وضبطها بحيث تنتج متجهات ذات فهم دقيق لمهمات دلالية بأطوال مختلفة.

[26] على الرابط: https://huggingface.co/sentence-transformers

[27] على الرابط: https://huggingface.co/Omartificial-Intelligence-Space

2.3. الباب الثالث في نماذج المرمزات الفاكّة

وهذه الهيكلية هي أول هيكلية صدرت للمحولات لمهمة الترجمة وتعتمد (كما في اسمها) على قسمين الأول يحاول فهم اللغة المترجمة منها في قسم المرمز والقسم الثاني الفاكّ (المفكك) فيحاول فهم اللغة المتَرجَمِ لها وربط ذلك بالمعلومات الآتية من المرمز المتعلقة باللغة الأصل ثم يتم تدريبه على توليد النصوص المترجمة (اللغة الثانية) تتابعيا.

جدول 4: النماذج المرمِّزة الفاكّة

الاسم	تاريخها	المصدر	عدد المعاملات	حجم البيانات[28]	معلومات المقسم	بيئة التدريب	مصدر البيانات	معلومات التدريب
AraT5 [22]	2021	جامعة بريتيش كولومبيا	220 مليون	29 مليار جزء لغوي	sentece Piece	خادم حوسبة جوجل	21.9 مليار جزء لغوي من تويتر و 7.1 مليار من الكتب والمقالات والشبكة (فصيح)	دُرب لمدة أربعين يومًا على كرت معالجة تنسرية TPU وقد صدر له نسخة ثانية أكبر[29]، وكذا تم ضبطه دقيقا على الترجمة باسم "ترجمان" [23]
nllb [24]	2020	شركة فيسبوك (ميتا)	-	-	-	خوادم فيسبوك	-	هذا نموذج ضخم جدا تم تدريبه على الترجمة في 200 لغة منها العربية
AraMUS [25]	2023	فريق الذكاء الادراكي Tonumus وشركة هواوي الحوسبية	13 مليار	530 جيجابايت	sentece Piece بعدد أجزاء 64,000	128 كرت شاشة A100	نصوص مسحوبة من الشبكة والمقالات	تم تدريب النموذج على 16 قطعة على كروت الشاشة لمدة شهرين باستعمال محسن Adafactor.

[28] يُذكر عدد الأجزاء اللغوية (بما ذكرته الورقة باستعمال مُقَسِّمهم) و إلا فيُذكر حجم البيانات.

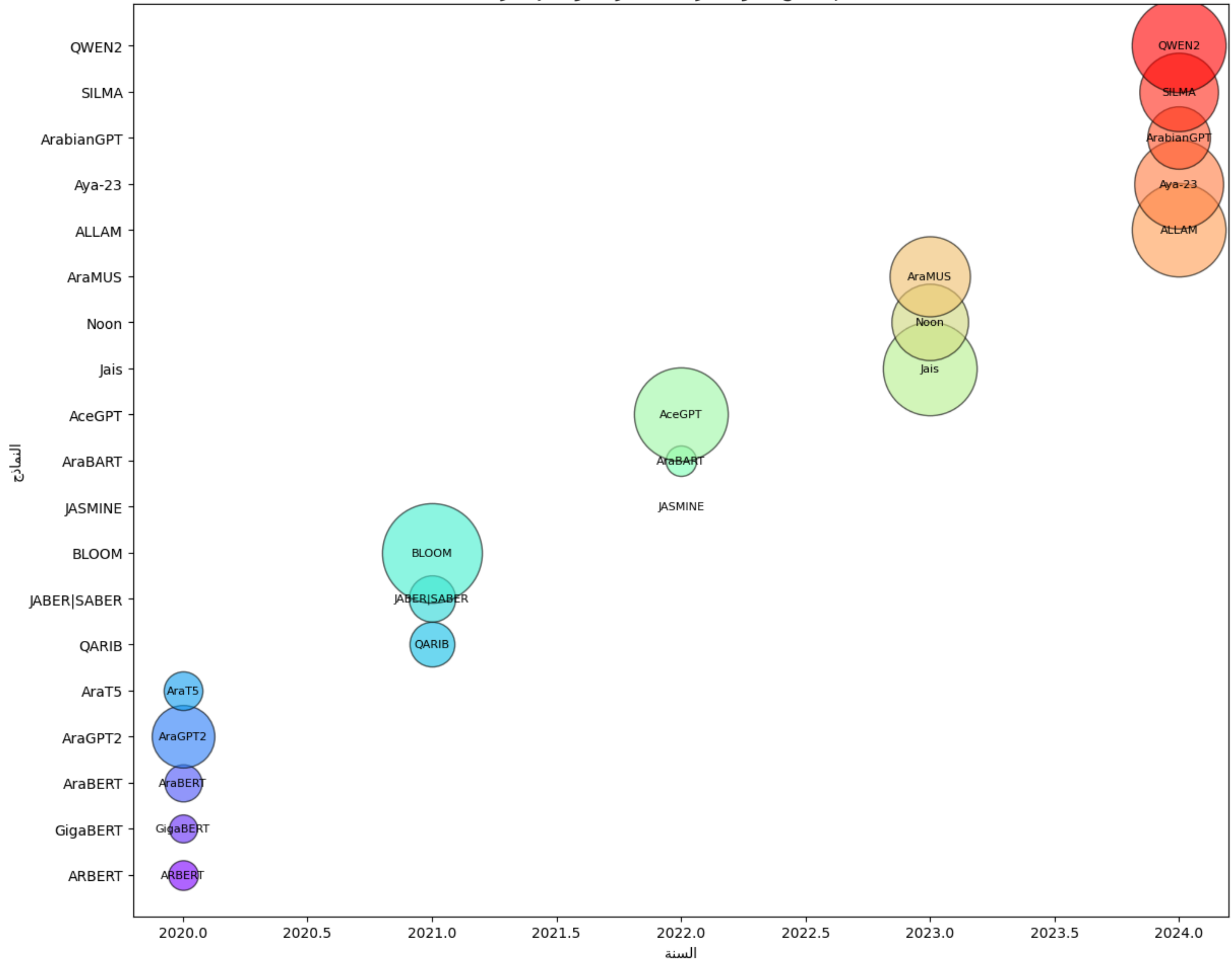
[29] لم يُصدر له ورقة و إنما نُشر فقط: https://huggingface.co/UBC-NLP/AraT5v2-base-1024

Peacock طاوس	طاوس (Peacock)	جامعة بريتيش كولومبيا و Invertible AI	-	-	-	-	نصوص وصور مركبة	ذُكر هذا النموذج لأنه يشابه هيكلية المرمز الفاك ولكن هنا الصورة تُربط بنظام محاثة عنها وهو الوحيد العربي للآن ولم يُنشر

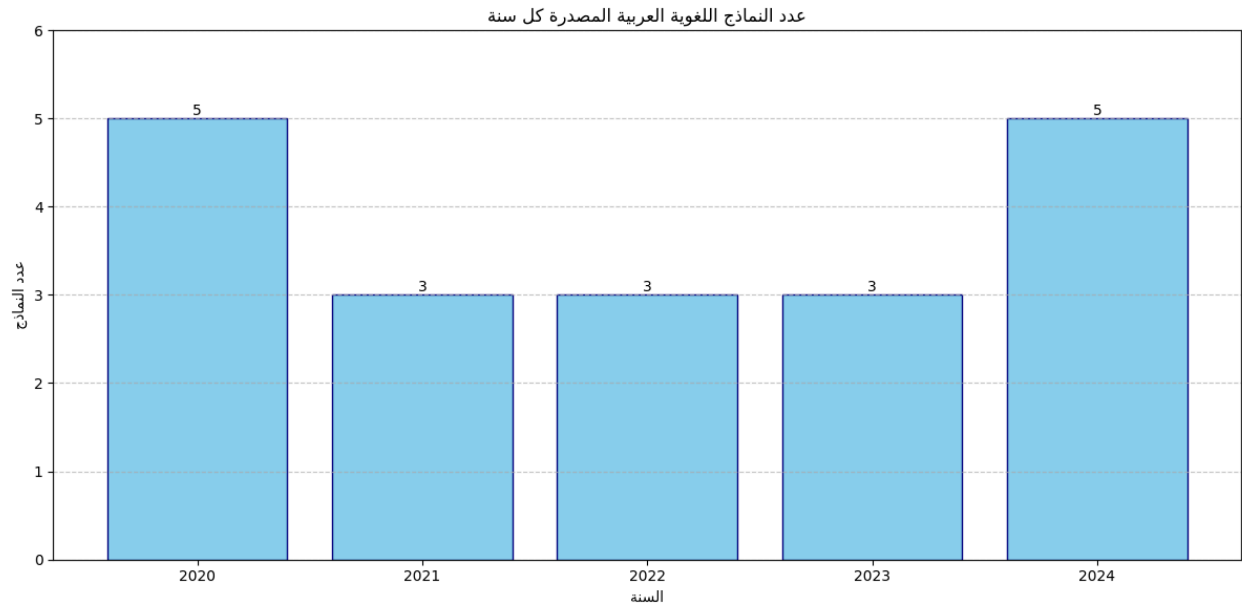
2.4. الرسوم البيانية

في هذا القسم سنذكر بعض الرسوم البيانية التي تساعد على استخراج بعض المعلومات المفيدة من مُعطيات النماذج وأرقامها.

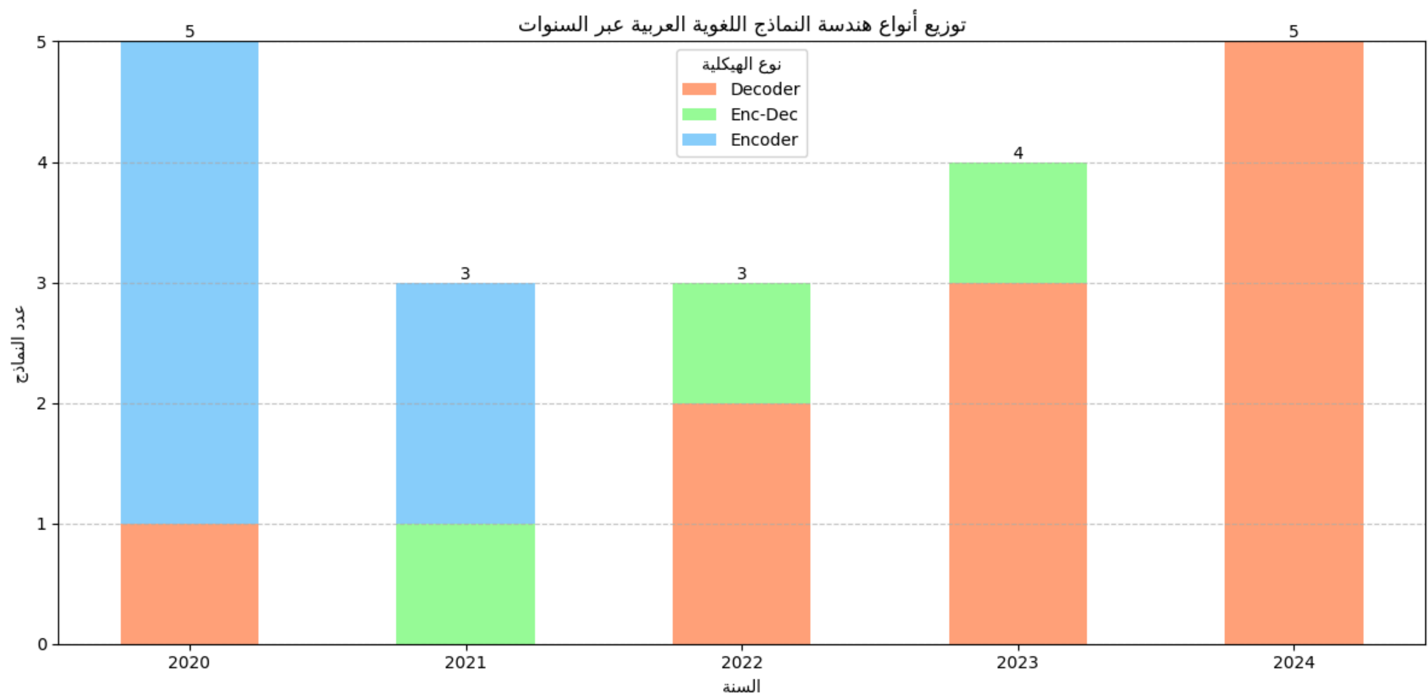
شكل1: عرض تاريخي مختصر للنماذج اللغوية مع سنة إصدارها (استُعملت النسخة الأساسية للنموذج)



شكل2: عدد النماذج (والعائلة تُعتبر واحدة هنا).

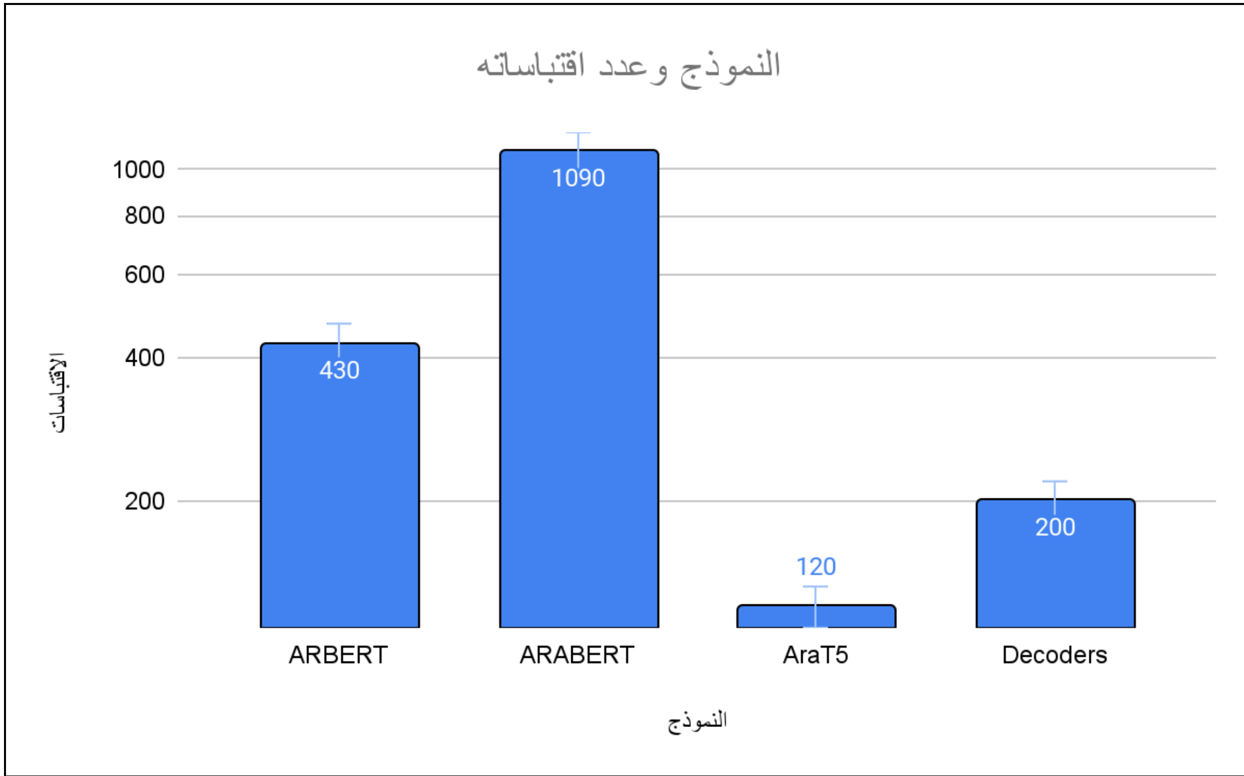


شكل3: عدد النماذج حسب هيكليَّتها كل سنة[30].



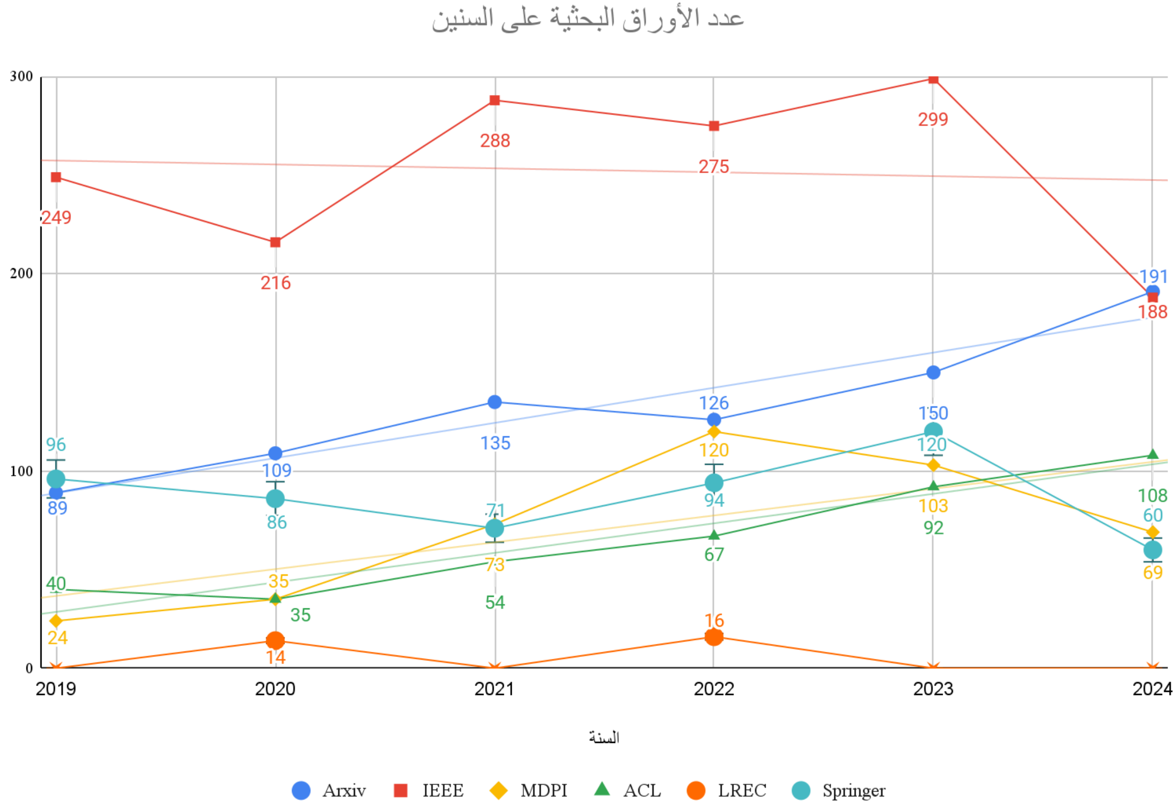
[30] يُشار إلى أن (Decoder) تعني المفكك و (Encoder) تعني مرمز (Enc-Dec) هي المرمز الفاك.

شكل4: عدد الاقتباسات من الأوراق البحثية[31]



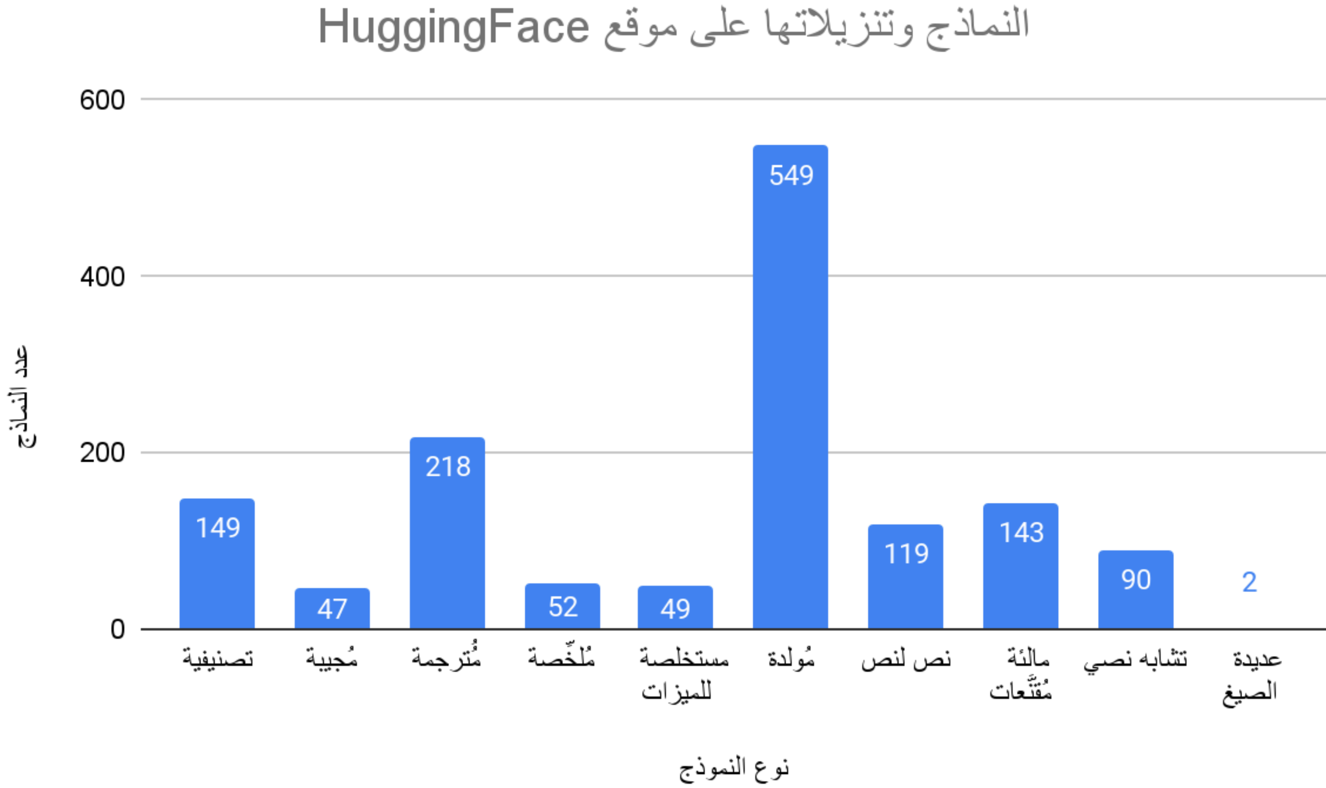
[31] قُمنا باختيار أشهر النماذج المفتوحة المصدر لأن الناس تستعملها وأما بالنسبة للنماذج المفككة فقمنا بجمع كل الاقتباسات في عمود واحد (وهي تشمل: AraGPT:110, Jais:55, Jasmine:3, Aya:25, Allam,ArabianGPT) نظرًا لقلة عددها إلا نموذج بلوم إذ هو متعدد اللغات وقد اقتُبست ورقته كثيرًا جدا بلغت 1500 ولكنها لا تعبر وصفيا عن العربية حصرًا.

شكل5: عدد الأوراق البحثية عن العربية ومعالجتها[32]



[32] قُمنا بسحب هذه الأرقام عن طريق البحث بكلمات مفتاحية عن العربية ونماذجها ومعالجتها. ويلاحظ بشكل واضح زيادة عدد الأوراق البحثية مع موجة النماذج العميقة من بعد عام 2019. ولم نستطع سحب عدد الأوراق على السنوات من موقع (ACL) لعدم دعمِ الموقع لخدمة البحث المتقدمة ولكن بعد بحث تقريبي فعليه ما يزيد على 10 آلاف ورقة بحثية عن العربية. والأرقام المعروض هي عدد الأوراق في المؤتمرات العربية فقط وهما مؤتمر (WANLP) وهو بين (2019-2022) ومؤتمرا (ArabicNLP) وهو الجديد بدلَ السابق وبدأ عام 2023. أما مؤتمر (LREC) فهو يقام كل سنتين في إيطاليا وتوقف عام 2022 ولهذا فله رقمان فقط. وننبه ختامًا إلى أن الأرقام قد يَدخُل فيها شيء من الدراسات التي لم تستعمل النماذج اللغوية بشكل مباشر لأن البحث بالتفصيل الدقيق متعسر في هذه المواقع خاصة عن اللغة العربية. ويُشار أيضا إلى بعض المجلات لم تُذكر مثل **Natural Language Engineering** لقلة عدد الأوراق (158 كليا) وعدم شُهرتها ولجعل الرسمة أوضح وأقل تعقيدا.

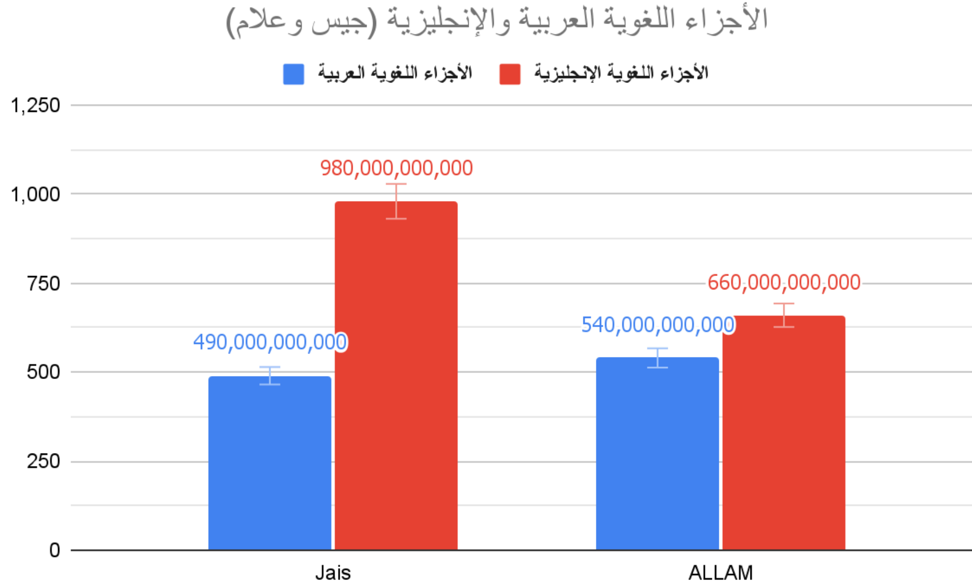
شكل6: أنواع النماذج وعددها على موقع HuggingFace[33]

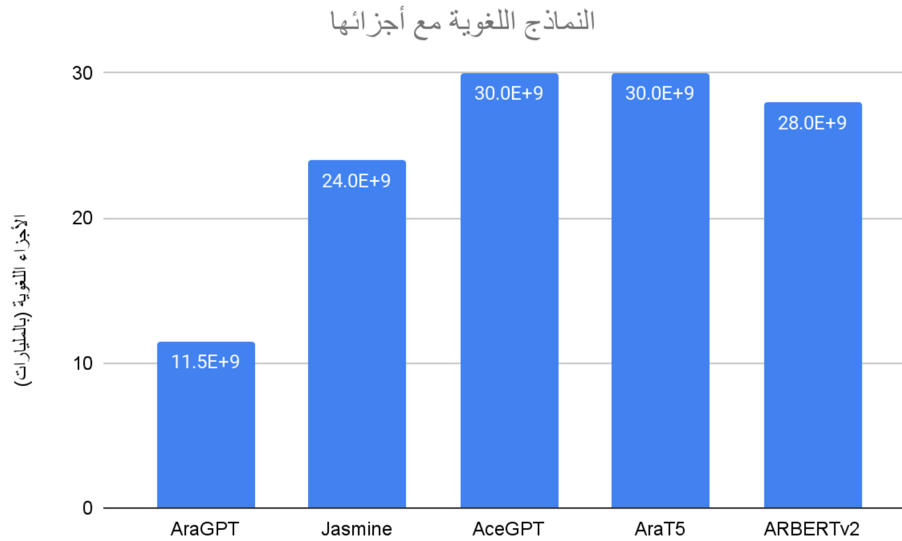


---

[33] بسبب تصنيف الموقع للنماذج فقد اضطررنا إلى اعتماد الأقسام التي على الموقع بدلا من ذكر عدد المفككات والمرمزات والمرمزات الفاكة. وهذا التقسيم المذكور لا شك أن فيه تداخلا بين هذه الأقسام الثلاثة. ومن الملاحظ كثرة النماذج المولدة بسبب الموجة الجديدة، ولكن فرضيَّتُنا أنها كثيرة لتجربتها خاصة أنه يُعلن عنها كل نموذج جديد بوقته.

شكل7: عدد الأجزاء اللغوية لنماذج[34]





3. أهمّ التطبيقات

3.1. تمهيد

في هذا القسم نعرض أشهر التطبيقات التي تطورت أو ظهرت مع ظهور النماذج اللغوية الكبيرة. إشارة مهمة أن هذه التطبيقات كلها يُمكن استخدام النماذج الضخمة المنشورة من قبل الشركات الكبيرة (ChatGPT) و (GEMINI) و (Claude) فهذه النماذج الكبيرة كلها

---

[34] هذه النماذج التي توفرت معلومات أجزاءها اللغوية (وقد ذُكرت في صورتين لفرق الحجم الكبير وأيضا لاستخدام الانجليزية فيها) وهناك نموذج (SABER|JABER) ومدونته حجمها: 115 جيجابايت ونموذج (ARAMUS) مدونته حجمها 530 جيجابايت.

استعملت سواء بواجهة أو بالإطار البرمجي (API) في هذه التطبيقات التي سنسردها تاليا وقد أشير في فصل التقييم ومعاييرها إلى الأوراق البحثية التي قيَّمت بعضًا منها وفي سردنا التالية سنذكر التطبيق ثُمَّ نشير إلى أحدث وأهم التحديثات التي استجدَّت مع اللغة العربية مما يفيد المستخدم العادي أو المطور أو الباحث إذ كلٌّ منهم يُمكنه استعماله بطريقة مخصصة.

3.2. أهم التطبيقات

- الترجمة

الترجمة من أهم المهمات المتعلقة باللغة العربية إذ فيها نقل المعرفة وخاصة مع قلة المصادر العلمية والمعرفية المكتوبة باللغة العربية فكان لا بُدَّ من تطوير هذا الجانب فصدرَت نماذج كبيرة متخصصة في الترجمة كما في "ترجمان" [23] وأيضا فهناك محاولات كثيرة لتقييم هذه النماذج في مهمة الترجمة كما مرَّ في فصل التقييم ومقاييسها وكذا رأينا أن الترجمة استُعملت في توسيع المدونات المُستعملة في تدريب النماذج اللغوية العربية كما في جيس [10] وعلام [16]. وأيضا مما كَثُر فيه البحث الآن (للأسف) ترجمة اللهجات والفصحى بعضهما من بعض ومن ذلك ما كان في مؤتمر اللغة العربية الثاني في تايلاند [27]. وكذلك فقد بدأت دراسات باللغة العربية تشير إلى هذا المجال [ع1] بأنه ما زال واعدًا ويحتاج تعديلات و إصلاحات كثيرة.

- الأسئلة والأجوبة (الإجابة)

هذه من أهم وأشهر تطبيقات النماذج اللغوية خاصة ذات الواجهات مثل (ChatGPT) وتشمل أمورًا كثيرة كطلب شرح مفهوم معين أو كتابة خطة لدراسة أو زيارة أو اقتراح مصادر أو أماكن. ولكن يعيب هذا التطبيق هلوسته (بأن تذكر وتكتب جوابا غير حقيقي أو غير صحيح) النماذج أحيانا ومن الأمور كذلك أن يُطلب من النموذج تيسير (تبسيط) نص أو مفهوم معين حتى يَسهل فهمه ومن الأوراق في ذلك [26].

- التلخيص

وهذه أيضا من أشهر تطبيقات النماذج اللغوية بحيثُ يُعطى النموذج نصا طويلا ويُطلب منه تلخيصه لآخر أصغر ويستخدمه بعض الطلبة في تلخيص الدروس والموظفين في تلخيص التقارير وغير ذلك.

- مهمات التصنيف وتشمل المشاعر وكشف المغاليط (مثل الأخبار) وتصنيف الأنواع (مثل أنواع المقالات أو اللهجات)

مهمات التصنيف كثيرة وأشهرها تصنيف المشاعر بأن يُعطى النموذج نصا معينا ليصنِّفه إلى أحد المشاعر الأساسية (عادة تكون سلبية أو إيجابية أو طبيعية محايدة) ومن ذلك [28] وكذلك من مهام التصنيف تصنيف النصوص إلى أنواعها الأساسية كتصنيف المقالات الإخبارية وكذلك تصنيف النصوص المغلوطة (الكاذبة). ومن الأمور الجديدة وهي تصنيف النصوص حسب جودتها لاستعمالها في تدريب النماذج اللغوية وكذا تصنيف النصوص حسب لهجتها[35].

- التعرف على الكيانات المُسمَّاة (NER)

وهذه من أقدم المهمات في المجال وهي في التعرف على الكيانات المُسماة وهي فئات محددة مسبقًا مثل الأسماء، والمنظمات أو المواقع أو الرموز الطبية أو التعبيرات الزمنية أو الكميات أو القيم النقدية أو النسب المئوية إلخ ومن آخر ما صدر ما كان في مؤتمر العربية الثاني [29].

- التشكيل النصي

وهذه مهمة من خصائص العربية وتهدف إلى "وضع علامات التشكيل الصحيحة على النصوص الخام" ومن أحسن النماذج الجديدة الصادرة نموذجا (CATT) (نموذج مرمز ونموذج مرمز مفكك) فقد أظهرا أداء عاليًا في التشكيل [30].

- تصحيح الأخطاء النصية واللغوية

تصحيح الأخطاء يُقصد به إصلاح النص المعطى (الحاوي عددا من الأخطاء سواءً كانت لغوي أو كتابية) و إرجاع نص صحيح وهذه الأخطاء إما أن تكون حقيقية أو موضوعة (مُصَنَّعة) لتدريب النموذج ومن الأوراق في هذا المجال مما استعمل (ChatGPT) في التصحيح [31] وبعضهم قام بتدريب وضبط نماذج محددة [32]

- استرداد المعلومات

وهذا المجال مما بزغ نجمه في المجال حديثا بما يُسمى "التوليد المُسترجِع" RAG بحيث ساعدت النماذج الجديدة على تشكيل تضمينات ذات جودة عالية ساعدت في فرز النصوص في متجهات محفوظة في قواعد معينة واسترجاع معلومات معينة حسب السؤال الموجه إلى النموذج وهنا مباحث تطوير نماذج عربية تنتج متجهات ذات جودة عالية لأنها مستخدمة في عدد كبير من التطبيقات. ومن الأوراق البحثية التي قامت بدراسة طُرق الاسترجاع مع العربية [33]. وكذا من المهام التي بُدءَ العمل عليها من السنة الماضية 2023 مهمة القاموس المعكوس بأن يُعطى التعريف ثُمَّ يُسترجع الكلمة المُعرفة (وهذا عكس عمل القاموس) وهذه ورقة المركز الأول في مؤتمر العربية الثاني [34].

---

[35] وهناك أيضا مهمة لقياس فصاحة النص ولكن مُخرجها رقم حقيقي

- كتابة النصوص البرمجية

وهذه أيضا مما ساعدت فيه النماذج أي كتابة النصوص البرمجية لعدد كبير من اللغات البرمجية خاصة بايثون واللغات البرمجية المشهورة ولكن في هذا التطبيق مبحث ضِعف القدرات المنطقية والبرمجية عند الطلاب وهذا مجال بحثي منفصل متعلق بمناهج وطرق التعليم وتغيّرها مع التطور البحثي.

- توليد النصوص المنوعة

هذا تطبيق واستخدام جديد ظهر مع النماذج اللغوية وهو توليد النصوص المتنوعة لتدريب النماذج اللغوية الكبيرة الأخرى وقد استعملته عدد من الشركات منها ميتا وجوجل ولكنَّ له أثرًا رجعيا سلبيا على المدى البعيد حسب بعض الأوراق البحثية [39,40]. وفي العربية فقد استُعملت هذه الطريقة في توليد عدد من مدونات النماذج السابق ذكرها إما للإختبار أو التدريب أو التقييم.

3.3. ملاحظة ختامية

وهناك أيضا بعض المجالات التي خُصِّصَت بالذكر لأهميَّتها وخطورتها وهي المجال الطبي والقانوني والشرعي وهذه أفردت ببعض الأوراق مما سبق ذكره في قسم التقييم من تقييم النماذج قانونيا وكذا في المجال الطبي [36] وأما في المجال الشرعي فهناك الكثير من الأبحاث المتعلقة بالقرآن فمنها مافي مؤتمر العربية عام 2023 من مهمة الأسئلة والأجوبة عن القرآن الكريم [37] وكذلك هناك دراسة دكتوراه مطولة لا تخلو من فوائد عن تطبيقات الذكاء الاصطناعي في القضاء [ع2] ودراسة ماجستير عن "توظيف تقنيات الذكاء الاصطناعي في الدعوة إلى الله" [ع3] ودراسة أخرى عن "توظيف الذكاء الاصطناعي في خدمة السنة النبوية" [ع4] وهناك بحوث عربية كثيرة جدا بدأت تظهر في الفترة الأخيرة فيما يتعلق بالذكاء الاصطناعي أخلاقيا وتعليميا وترجميا [ع5]. ومن المجالات أيضا التي لم تشتهر ولكن الجهود فيها كبيرة مجال تخصيص النماذج في مجال معين (إما بالضبط الدقيق أو بالتدريب من الصفر) كالبنوك والشركات عموما، خاصة في مهمات استخراج المعلومات وتحويلها من الهيئة غير المُهيكلة إلى صيغة مُهيكلة سهلٌ التعامل معها (كالجداول) وأيضا مهمات التلخيص كما في مجال التداول والتقارير المالية (توليدا أو استفادة).

وأشير ختاما إلى مجال ظهر مع النماذج اللغوية وهو "هندسة الأوامر" وخلاصته في كتابة الأوامر المناسبة التي تحصل بها على أحسن الردود من النماذج وله دراسات كثيرة ومما صدر عن العربية هذه الورقة [35].

ولكثرة التطبيقات والأبحاث في خدمة العربية فإنه يصعب حصرها بل يعسر الإشارة لكل جهد فيها فجدير بالذكر مجاميع الأوراق البحثية التي تصدر عن المؤتمرات الخاصة باللغة العربية مثل [38].

4. التحديات والتوجهات المُستقبلية

سنحاول في هذا القسم أن نحصر التحديات والتوجهات المُستخلصة من المعلومات المجموعة في القسمين السابقين (حصر النماذج والتطبيقات)

1. التحديات

- نقص في المدونات النصية العربية الصِّرفَة وهذا أمر ملاحظ من كل الأوراق البحثية السابقة إذ دائما يُشار لها في قسم "بيانات التدريب" خاصة في النماذج الضخمة مثل "جيس (Jais)" أو "علام(ALLAM)" نسبة العربية دائما أقل من الإنجليزية.
- قلة العاملين العرب في المجال ويتبيَّن هذا من رؤية أسماء الباحثين في الأوراق البحثية وخاصة في النماذج الضخمة الجديدة مثل "جيس (Jais)" أو "علام (ALLAM)" أو "أيس Ace" وهذه تمَّ تدريبها إما على أرض عربية بأيدٍ غير عربية مجملا أو بالشراكة مع جهة عربية فضلًا عن النماذج التي درَّبتها جهات أجنبية مثل "آية Aya" أو "كَوِين QWEN" والتي تدعم العربية بشكل جيد جدا
- ضعف القدرات الحوسبية في تدريب النماذج العربية وهذا نسبية طبعا من جهة لأخرى ولكن مقارنة مع الشركات الخاصة فإن الأَعتُد (جمع عتاد) ضعيفة وقليلة حتى في النماذج الضخمة مثل "علام (ALLAM)" أو "جيس (Jais)" (والذي دُرِّبَ على حاسوب فائق -خارق-) ببون شاسع مثلا مع بُنية شركة "ميتا" والتي ستبلغ (حسب المُخطط له سنة 2024) 350,000 ألفَ كرت شاشة من نوع (H100).
- تشتت الجهود العربية من جميع الدول العربية من إنتاج وتطوير بدون إطار إداري أو توجه منهجي.
- ندرة المحتوى التعليمي الجادّ باللسان العربي مكتوبًا أو مرئيًّا الخاص بشرح ما يستجد في مجال معالجة اللغة العربية عموما والنماذج اللغوية خصوصا.
- انعدام النماذج العربية اللغوية عديدة الصيغ (Multimodal) إذ إلى أثناء كتابة هذه الورقة لم تُنشر إلا ورقة واحدة ونموذج واحد على موقع (HuggingFace).
- استعمال المقسمات المشهورة في تدريب النماذج اللغوية كمقسم BPE أو Sentence Piece والخوارزمية المُعتمدة في هذه المقسمات فيها نوع اختلاف (تعارض) مع طبيعة صرف اللغة العربية (والتي فيها زوائد في كل مواضع الكلمة) وكذا طابع كتابة اللغة العربية فيما يتعلق بالتشكيل والتمطيط وغيرها، ودراسة تأثيرها على تدريب النموذج وأداءه.
- صب جهود كبيرة في نماذج المفككات بسبب موجة السنوات الماضية مع أن الأرقام تدل على كثرة استخدام المرمزات بشكل أكبر سواءً في اقتباسات الورقة أو تنزيلات النماذج.
- النماذج العملاقة مثل علام وجيس تُعتبر نماذج انجليزية إذ نسبتها فيها أكبر، حتى النصوص المترجمة فيها نسبتها كبيرة جدا (في علَّام بلغت النصف) فهذا يُوثر على الفهم الثقافي للنموذج بل وحتى الفهم اللغوي.

2. التوجهات المستقبلية

- هناك اهتمام بالنماذج المرمزة (Encoder) ويتبين هذا من كثرة اقتباسات الباحثين وذكرها في عدد من البحوث التطبيقية مع أن هذه الهيكلية من النماذج هناك شبه توقف في تدريب جديدٍ منها من بعد 2021، وهذا يعيد فتح الأبواب لتطوير وتحسين البحث فيها، وأيضًا يدخل تحت هذا المبحث، مبحث "التشابه الدلالي(Sentence Similarly)" أو التشابه الجُمَلي وهو يعتمد على هذه الهيكلية غالبًا.
- زيادة الجهود والأبحاث لتوسيع المدونات العربية من خلال تطوير أساليب استخراج النصوص ذات الجودة العالية وكذا توحيد الجهود في نشر المدونات الضخمة وثانيًا تطوير أساليب استخراج النصوص من المصورات لأن هناك عددًا كبيرا من الكتب التي لم تُرَقمن بعد، وأخيرا تحسين الترجمة سواءً في كثرة الإنتاج البشري أو في تنمية النماذج الآلية.
- تكريس الجهود في إثراء المحتوى التعليمي التقني الدقيق المُحَقَّق ليساعد الطالب والباحث العربي على الوصول للمعلومة الصحيحة.
- تدريب العاملين العرب في المجال فإنه لا بد من أفضلية من وجود الباحث العربي في بحوث تطوير النماذج سواءً في اختيار المدونات أو في تقييم المخرجات فيما يتعلق بالإدراك العام للثقافة العربية.
- توحيد الجهود في تأطير الجهود العربية من جميع البلدان العربية (وغيرها من البلدان المهتمة) في كل ما يتعلق بالمجال هذا خصوصًا وما يخص العربية عمومًا.
- التركيز على خدمة العربية الفُصحى وجعل الجهود المتعلقة باللهجة تنصب في ردِّها للفصيحة لا لذاتها.
- دراسة وتطوير المقسمات اللغوية بما يتداخل مع اللغة العربية وطبيعتها الصرفية والكتابية.

5. الخاتمة

قدم هذا الفصل نظرة شاملة لتطور النماذج اللغوية الكبيرة، مع تركيز خاص على النماذج العربية منذ عام **2020**. استعرضنا المراحل التاريخية الرئيسية، بدءًا من الشبكات العصبية وصولاً إلى نماذج المحولات، مسلطين الضوء على التنافس والتعاون بين الشركات والمؤسسات البحثية. كما ناقشنا بالتفصيل النماذج العربية الضخمة المفككة، محللين مدونات التدريب، وحجم النماذج، وطرق التقييم. وأخيرا، استعرضنا النماذج ذات هيكلية المرمز والمرمز المفكك، مبرزين أهم تطبيقاتها الحالية والناشئة.

رغم التقدم الملحوظ، تواجه النماذج العربية تحديات جوهرية تشمل التحيز اللغوي والثقافي، خاصة في تمثيل اللهجات المحلية والسياقات الثقافية المتنوعة. كما تعاني من محدودية البيانات عالية الجودة والتنوع، والحاجة لدمج معرفة متخصصة في مجالات محددة. إضافة إلى ذلك، تبرز ضرورة معالجة القضايا الأخلاقية والقانونية المتعلقة بالخصوصية وحماية البيانات.

مستقبل النماذج اللغوية العربية الكبيرة يبشر بتطورات مثيرة، منها تطوير نماذج متعددة اللغات لتعزيز الترجمة وفهم السياق عبر الثقافات، وتحسين معالجة اللهجات العربية المختلفة بدقة أعلى. كما نتوقع دمج معرفة أعمق بالتراث العربي والإسلامي لتحسين الأداء في السياقات الثقافية والتاريخية، وتوسيع نطاق التطبيقات لتشمل التعليم الإلكتروني، الرعاية الصحية عن بعد، والخدمات الحكومية الرقمية.

تمثل هذه النماذج فرصة هائلة لتعزيز الابتكار والتواصل في العالم العربي، لكن يجب مواجهة التحديات الأخلاقية والتقنية بحكمة لضمان تأثير إيجابي على المجتمع. من المهم تطوير هذه النماذج بشكل يراعي الخصوصية الثقافية واللغوية للعالم العربي، مع الحفاظ على التنوع اللهجي والثقافي. كما يجب الاهتمام بتطوير موارد لغوية عربية عالية الجودة لتحسين أداء هذه النماذج.

أخيرًا، يعد التعاون بين المؤسسات الأكاديمية والصناعية في العالم العربي أمرًا ضروريا لتسريع تطوير هذه التقنيات وتطبيقها في مجالات تعود بالنفع على المجتمع العربي. مع استمرار البحث والتطوير، نتوقع أن تلعب النماذج اللغوية العربية الكبيرة دورًا محوريًا في تشكيل مستقبل التقنية والمعرفة في المنطقة العربية، مع ضرورة الموازنة بين الابتكار والقيم الثقافية والأخلاقية.

6. المراجع العربية

[ع1] صونيا أسمهان حليمي. "الترجمة الآلية وأنموذج الترجمة إلى العربية: المشهد الراهن." مجلّة الإيسيسكو للُّغة العربيَّة 1, no. 1 (2024): 295-313.

[ع2] الجلعود، أروى. أحكام تطبيقات الذكاء الاصطناعي في القضاء. جمعية قضاء، 2024. (أصلها رسالة دكتوراه)

[ع3] الحربي، ابتسام بنت عبد الله. توظيف تقنيات الذكاء الاصطناعي في الدعوة إلى الله. المعهد العالي للدعوة والاحتساب، جامعة الإمام محمد بن سعود الإسلامية، 2019. (رسالة ماجستير)

[ع4] آل مُعَدِّي, محمد بن عبد الله, and محمد بن عبد الله. "توظيف الذكاء الاصطناعي في خدمة السنة النبوية-رؤية في أبرز المخاطر والتحديات." الدراية 24, no. 24 (2024): 161-182.

[ع5] المركز العربي الديمقراطي في برلين له عدد من البحوث والكتب المنشورة الجديدة في المجال وكذا هناك عدد كبير من دراسات الدكتوراه والماجستير وقد أشرنا لبعضها وهناك عدد من الكتب المنشورة مثل كتب الدكتور "طعيمة" وكذا ما ينشره المجمع من دراسات لسانية وكتب (ككتب مركز الملك عبد الله قديما) وكذا منشورات الجامعات مثل سعود والإمام ولعل هذا انطلاقة لحصر الجهود العربية في الذكاء الاصطناعي وترتيبها

[ع6] الهيئة السعودية للبيانات والذكاء الاصطناعي. معجم البيانات والذكاء الاصطناعي. د. ن, 2022.

[ع7] شركة صخر ومعلوماتها الأساسية على الموسوعة الحرة Wikipedia
https://ar.wikipedia.org/wiki/%D8%B5%D8%AE%D8%B1_(%D8%B4%D8%B1%D9%83%D8%A9)

[ع8] ملخص محاضرات في اللسانيات التطبيقية للدكتور مصطفى طويل (منشور على الشبكة)
https://moodle.univ-chlef.dz/ar/course/info.php?id=1426

7. المراجع الأجنبية

[1] Antoun, Wissam, Fady Baly, and Hazem Hajj. "AraGPT2: Pre-trained transformer for Arabic language generation." arXiv preprint arXiv:2012.15520 (2020).

[2] Workshop, BigScience, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana IliĆ, Daniel Hesslow et al. "Bloom: A 176b-parameter open-access multilingual language model." arXiv preprint arXiv:2211.05100 (2022).

[3] Laurençon, Hugo, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra et al. "The bigscience roots corpus: A 1.6 tb composite multilingual dataset." Advances in Neural Information Processing Systems 35 (2022): 31809-31826.

[4] Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur et al. "The llama 3 herd of models." arXiv preprint arXiv:2407.21783 (2024).

[5] Abdul-Mageed, Muhammad, Abdelrahim Elmadany, Alcides Inciarte, and Md Tawkat Islam Khondaker. "JASMINE: Arabic GPT Models for Few-Shot Learning." In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 16721-16744. 2023.

[6] Huang, Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen et al. "AceGPT, Localizing Large Language Models in Arabic." arXiv preprint arXiv:2309.12053 (2023).

[7] Zhu, Jianqing, Huang Huang, Zhihang Lin, Juhao Liang, Zhengyang Tang, Khalid Almubarak, Abdulmohsen Alharthi et al. "Second Language (Arabic) Acquisition of LLMs via Progressive Vocabulary Expansion."

[8] Seelawi, Haitham, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. "ALUE: Arabic language understanding evaluation." In Proceedings of the Sixth Arabic Natural Language Processing Workshop, pp. 173-184. 2021.

[9] Hellsten, Simon. "Incremental Re-tokenization in BPE-trained SentencePiece Models." (2024).

[10] Sengupta, Neha, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall et al. "Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models." arXiv preprint arXiv:2308.16149 (2023).

[11] Hugging face page about new Jais models (include the new report) https://huggingface.co/inceptionai/jais-adapted-70b (has info about all)

[12] Aryabumi, Viraat, John Dang, Dwarak Talupuru, Saurabh Dash, David Cairuz, Hangyu Lin, Bharat Venkitesh et al. "Aya 23: Open weight releases to further multilingual progress." arXiv preprint arXiv:2405.15032 (2024).

[13] Singh, Shivalika, Freddie Vargus, Daniel Dsouza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya et al. "Aya dataset: An open-access collection for multilingual instruction tuning." arXiv preprint arXiv:2402.06619 (2024).

[14] Üstün, Ahmet, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari et al. "Aya model: An instruction finetuned open-access multilingual language model." arXiv preprint arXiv:2402.07827 (2024).

[15] Koubaa, Anis, Adel Ammar, Lahouari Ghouti, Omer Necar, and Serry Sibaee. "ArabianGPT: Native Arabic GPT-based Large Language Model." (2024).

[16] Saiful Bari, M., Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza et al. "ALLaM: Large Language Models for Arabic and English." arXiv e-prints (2024): arXiv-2407.

[17] Yang, An, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li et al. "Qwen2 technical report." arXiv preprint arXiv:2407.10671 (2024).

[18] Antoun, Wissam, Fady Baly, and Hazem Hajj. "Arabert: Transformer-based model for arabic language understanding." arXiv preprint arXiv:2003.00104 (2020).

[19] Abdul-Mageed, Muhammad, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. "ARBERT & MARBERT: Deep bidirectional transformers for Arabic." arXiv preprint arXiv:2101.01785 (2020).

[20] Lan, Wuwei, Yang Chen, Wei Xu, and Alan Ritter. "An empirical study of pre-trained transformers for Arabic information extraction." arXiv preprint arXiv:2004.14519 (2020).

[21] Ghaddar, Abbas, Yimeng Wu, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao Xing, Yasheng Wang et al. "Jaber and saber: Junior and senior arabic bert." arXiv preprint arXiv:2112.04329 (2021).

[22] Nagoudi, El Moatez Billah, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. "AraT5: Text-to-text transformers for Arabic language generation." arXiv preprint arXiv:2109.12068 (2021).

[23] Nagoudi, El Moatez Billah, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. "TURJUMAN: A public toolkit for neural Arabic machine translation." arXiv preprint arXiv:2206.03933 (2022).

[24] Costa-jussà, Marta R., James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi et al. "No language left behind: Scaling human-centered machine translation." arXiv preprint arXiv:2207.04672 (2022).

[25] Alghamdi, Asaad, Xinyu Duan, Wei Jiang, Zhenhai Wang, Yimeng Wu, Qingrong Xia, Zhefeng Wang et al. "Aramus: Pushing the limits of data and model scale for arabic natural language processing." arXiv preprint arXiv:2306.06800 (2023).

[26] SalahEldin, Yousef, and Caroline Sabty. "Simplify: Automatic Arabic Sentence Simplification using Word Embeddings." In Proceedings of ArabicNLP 2023, pp. 418-422. 2023.

[27] Abdul-Mageed, Muhammad, Amr Keleg, AbdelRahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. "NADI 2024: The Fifth Nuanced Arabic Dialect Identification Shared Task." arXiv preprint arXiv:2407.04910 (2024).

[28] Al-Thubaity, Abdulmohsen, Sakhar Alkhereyf, Hanan Murayshid, Nouf Alshalawi, Maha Omirah, Raghad Alateeq, Rawabi Almutairi, Razan Alsuwailem, Manal Alhassoun, and Imaan Alkhanen. "Evaluating ChatGPT and bard AI on Arabic sentiment analysis." In Proceedings of ArabicNLP 2023, pp. 335-349. 2023.

[29] Jarrar, Mustafa, Nagham Hamad, Mohammed Khalilia, Bashar Talafha, AbdelRahim Elmadany, and Muhammad Abdul-Mageed. "WojoodNER 2024: The Second Arabic Named Entity Recognition Shared Task." arXiv preprint arXiv:2407.09936 (2024).

[30] Alasmary, Faris, Orjuwan Zaafarani, and Ahmad Ghannam. "CATT: Character-based Arabic Tashkeel Transformer." arXiv preprint arXiv:2407.03236 (2024).

[31] Kwon, Sang Yun, Gagan Bhatia, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. "Beyond English: Evaluating LLMs for Arabic Grammatical Error Correction." arXiv preprint arXiv:2312.08400 (2023).

[32] Salhab, Mahmoud, and Faisal Abu-Khzam. "Araspell: A deep learning approach for arabic spelling correction." arXiv preprint arXiv:2405.06981 (2024).

[33] El-Beltagy, Samhaa R., and Mohamed A. Abdallah. "Exploring Retrieval Augmented Generation in Arabic." arXiv preprint arXiv:2408.07425 (2024).

[34] Sibaee, Serry, Abdullah Alharbi, Samar Ahmad, Omer Nacar, Anis Koubaa, and Lahouari Ghouti. "ASOS at KSAA-CAD 2024: One Embedding is All You Need for Your Dictionary." In Proceedings of The Second Arabic Natural Language Processing Conference, pp. 697-703. 2024.

[35] El-Sheikh, Abdelrahman, Ahmed Elmogtaba, Kareem Darwish, Muhammad Elmallah, Ashraf Elneima, and Hassan Sawaf. "Creating Arabic LLM Prompts at Scale." arXiv preprint arXiv:2408.05882 (2024).

[36] ALMutairi, Mariam, Lulwah AlKulaib, Melike Aktas, Sara Alsalamah, and Chang-Tien Lu. "Synthetic Arabic Medical Dialogues Using Advanced Multi-Agent LLM Techniques." In Proceedings of The Second Arabic Natural Language Processing Conference, pp. 11-26. 2024.

[37] Malhas, Rana, Watheq Mansour, and Tamer Elsayed. "Qur'an QA 2023 Shared Task: Overview of Passage Retrieval and Reading Comprehension Tasks over the Holy Qur'an." Proceedings of ArabicNLP 2023 (2023): 690-701.

[38] مجموع الأوراق البحثية في مؤتمري اللغة العربية الأول في سنغافورة والثاني في تايلند مما يحسن مراجع لكثرة أوراقه وتطبيقات وأفكاره وكذا من المواقع المفيدة arxiv ففيه تنشر مسودات مشاريع كثيرة فابحث فقد عن Arabic في صندوق البحث واختر العناوين.

- Habash, Nizar, Houda Bouamor, Ramy Eskander, Nadi Tomeh, Ibrahim Abu Farha, Ahmed Abdelali, Samia Touileb et al. "Proceedings of The Second Arabic Natural Language Processing Conference." In Proceedings of The Second Arabic Natural Language Processing Conference. 2024.
- Sawaf, Hassan, Samhaa R. El-Beltagy, Wajdi Zaghouani, Walid Magdy, Ahmed Abdelali, Nadi Tomeh, Ibrahim Abu Farha et al. "Proceedings of ArabicNLP 2023." In Proceedings of ArabicNLP 2023. 2023.

[39] Shumailov, Ilia, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. "AI models collapse when trained on recursively generated data." Nature 631, no. 8022 (2024): 755-759.

[40] Long, Lin, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. "On llms-driven synthetic data generation, curation, and evaluation: A survey." arXiv preprint arXiv:2406.15126 (2024).

[41] Chang, Yupeng, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen et al. "A survey on evaluation of large language models." ACM Transactions on Intelligent Systems and Technology 15, no. 3 (2024): 1-45.

[42] Hu, Taojun, and Xiao-Hua Zhou. "Unveiling LLM Evaluation Focused on Metrics: Challenges and Solutions." arXiv preprint arXiv:2404.09135 (2024).

[43] Khondaker, Md Tawkat Islam, Abdul Waheed, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. "Gptaraeval: A comprehensive evaluation of chatgpt on Arabic nlp." arXiv preprint arXiv:2305.14976 (2023).

[44] Alyafeai, Zaid, Maged S. Alshaibani, Badr AlKhamissi, Hamzah Luqman, Ebrahim Alareqi, and Ali Fadel. "Taqyim: Evaluating Arabic nlp tasks using chatgpt models." arXiv preprint arXiv:2306.16322 (2023).

[45] Alghamdi, Emad A., Reem I. Masoud, Deema Alnuhait, Afnan Y. Alomairi, Ahmed Ashraf, and Mohamed Zaytoon. "AraTrust: An Evaluation of Trustworthiness for LLMs in Arabic." arXiv preprint arXiv:2403.09017 (2024).

[46] Abdelali, Ahmed, Hamdy Mubarak, Shammur Absar Chowdhury, Maram Hasanain, Basel Mousi, Sabri Boughorbel, Yassine El Kheir et al. "LAraBench: Benchmarking Arabic AI with Large Language Models." arXiv preprint arXiv:2305.14982 (2023).

[47] Almazrouei, Ebtesam, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo et al. "AlGhafa Evaluation Benchmark for Arabic Language Models." In Proceedings of ArabicNLP 2023, pp. 244-275. 2023.

[48] Elmadany, AbdelRahim, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. "ORCA: A challenging benchmark for Arabic language understanding." arXiv preprint arXiv:2212.10758 (2022).

[49] Seelawi, Haitham, Ibraheem Tuffaha, Mahmoud Gzawi, Wael Farhan, Bashar Talafha, Riham Badawi, Zyad Sober, Oday Al-Dweik, Abed Alhakim Freihat, and Hussein Al-Natsheh. "ALUE: Arabic language understanding evaluation." In Proceedings of the Sixth Arabic Natural Language Processing Workshop, pp. 173-184. 2021.

[50] Elmadany, Abdelrahim, Ahmed El-Shangiti, and Muhammad Abdul-Mageed. "Dolphin: A Challenging and Diverse Benchmark for Arabic NLG." In Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 1404-1422. 2023.

[51] Atef, Adel, Bassam Mattar, Sandra Sherif, Eman Elrefai, and Marwan Torki. "AQAD: 17,000+ arabic questions for machine comprehension of text." In 2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA), pp. 1-6. IEEE, 2020.

[52] Abdallah, Abdelrahman, Mahmoud Kasem, Mahmoud Abdalla, Mohamed Mahmoud, Mohamed Elkasaby, Yasser Elbendary, and Adam Jatowt. "Arabicaqa: A comprehensive dataset for arabic question answering." In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 2049-2059. 2024.

[53] لا يوجد لمدونة بلسم ورقة بحثية و إنما نُشر على موقعهم بضع معلومات عنها ومنها جوانب الاختبار وقد رفعوا أمثلة تطوير من كل مهمة https://benchmarks.ksaa.gov.sa/b/balsam

[54] Koto, Fajri, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak et al. "Arabicmmlu: Assessing massive multitask language understanding in arabic." arXiv preprint arXiv:2402.12840 (2024).

[55] Alghamdi, Reem, Zhenwen Liang, and Xiangliang Zhang. "Armath: a dataset for solving arabic math word problems." (2022).

[56] Hijazi, Faris, Somayah AlHarbi, Abdulaziz AlHussein, Harethah Abu Shairah, Reem AlZahrani, Hebah AlShamlan, Omar Knio, and George Turkiyyah. "ArabLegalEval: A Multitask Benchmark for Assessing Arabic Legal Knowledge in Large Language Models." arXiv preprint arXiv:2408.07983 (2024).

[57] Naous, Tarek, Michael J. Ryan, Alan Ritter, and Wei Xu. "Having beer after prayer? measuring cultural bias in large language models." arXiv preprint arXiv:2305.14456 (2023).

[58] Siddhant, Aditya, Junjie Hu, Melvin Johnson, Orhan Firat, and Sebastian Ruder. "Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization." In Proceedings of the International Conference on Machine Learning 2020, pp. 4411-4421. 2020.

[59] Soliman, Abu Bakr, Kareem Eissa, and Samhaa R. El-Beltagy. "Aravec: A set of arabic word embedding models for use in arabic nlp." Procedia Computer Science 117 (2017): 256-265.

[60] Sabri, Tarik, Said Bahassine, Omar El Beggar, and Mohamed Kissi. "An improved Arabic text classification method using word embedding." International Journal of Electrical & Computer Engineering (2088-8708) 14, no. 1 (2024).

[61] Lo, Andrew W., and Manish Singh. "From ELIZA to ChatGPT: The Evolution of NLP and Financial Applications." (2023).

[62] Darwish, Kareem, Nizar Habash, Mourad Abbas, Hend Al-Khalifa, Huseein T. Al-Natsheh, Houda Bouamor, Karim Bouzoubaa et al. "A panoramic survey of natural language processing in the Arab world." Communications of the ACM 64, no. 4 (2021): 72-81.

[63] Abdelali, Ahmed, Sabit Hassan, Hamdy Mubarak, Kareem Darwish, and Younes Samih. "Pre-training bert on arabic tweets: Practical considerations." arXiv preprint arXiv:2102.10684 (2021).

[64] Alyafeai, Zaid, Maged S. Al-shaibani, Mustafa Ghaleb, and Irfan Ahmad. "Evaluating various tokenizers for Arabic text classification." Neural Processing Letters 55, no. 3 (2023): 2911-2933.

[65] Alwajih, Fakhraddin, El Moatez Billah Nagoudi, Gagan Bhatia, Abdelrahman Mohamed, and Muhammad Abdul-Mageed. "Peacock: A Family of Arabic Multimodal Large Language Models and Benchmarks." arXiv preprint arXiv:2403.01031 (2024).

[66] Nacar, Omer, and Anis Koubaa. "Enhancing Semantic Similarity Understanding in Arabic NLP with Nested Embedding Learning." arXiv preprint arXiv:2407.21139 (2024).

[67] Qarah, Faisal, and Tawfeeq Alsanoosy. 2024. "A Comprehensive Analysis of Various Tokenizers for Arabic Large Language Models" Applied Sciences 14, no. 13: 5696. https://doi.org/10.3390/app14135696